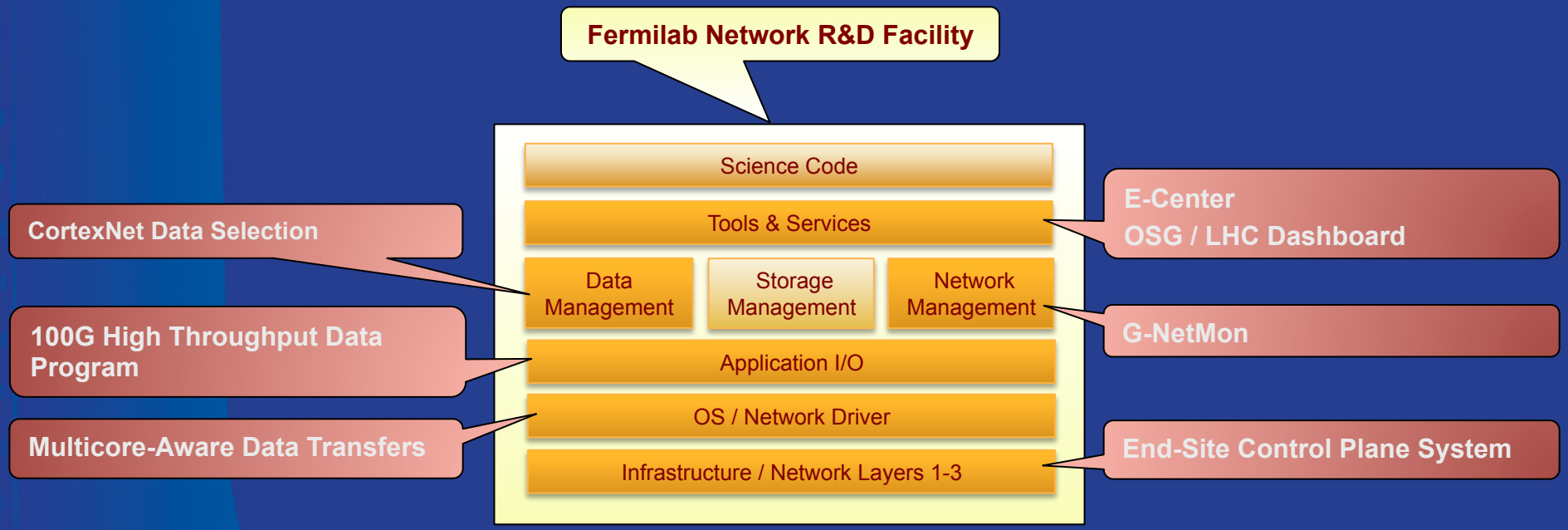
High Throughput Data Program (HTDP)

Parag Mhashilkar
Grid and Cloud Computing Department, Computing Sector
Fermi National Accelerator Laboratory

SCD Project Meeting - June 19, 2013

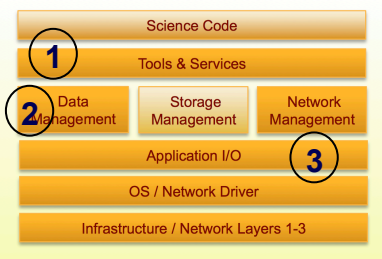
Network R&D at Fermilab



- A diverse program of work that spans all layers of computing for scientific discovery
- A collaborative process benefitting from the effort of multiple research organizations
- A broad range of activities internally and externally funded

R&D effort in HTDP

Three main thrusts:

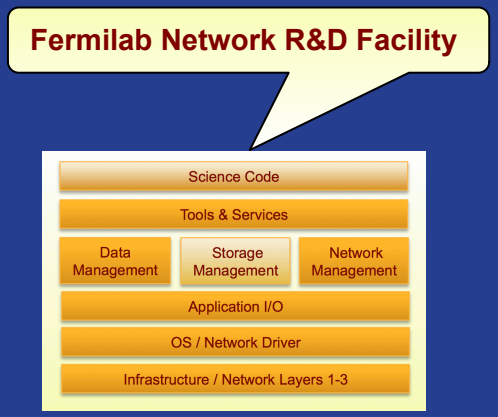


1. Providing tools & services to enable users / applications to optimize use of the network
   - Collaborating with the OSG Network Area for the deployment of PerfSONAR at 100 OSG facilities
   - Aggregating and displaying data through E-Center and the OSG Dashboard for end-to-end hop-by-hop paths across network domains

   Proposed integration with Data Management through network-aware data source selection – CortexNET
3. Application-level R&D through the High Throughput Data Program
   - R&D on 100G for production use by CMS & FNAL high-capacity high-throughput Storage facility
   - Identifying gaps in data movement middleware for the applications used for scientific discovery
     - GridFTP, SRM, Globus Online, XRootD, Frontier / Squid, NFS v4
     - List driven by stakeholder requests

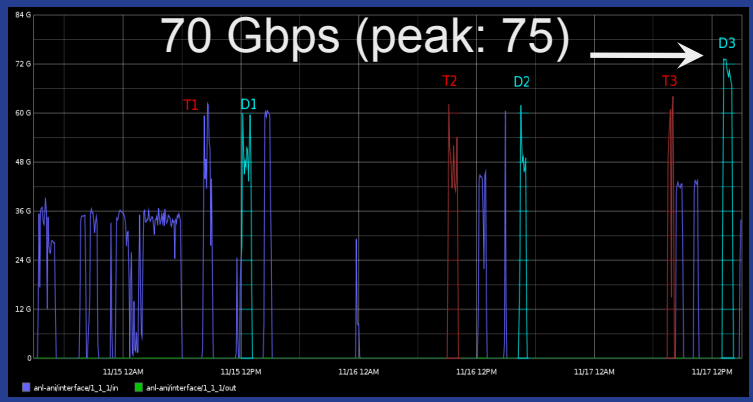
Goals of 100G Program at Fermilab

- Experiment analysis systems include a deep stack of software layers and services.
- Need to ensure these are effectively functional at the 100G scale end-to-end.
  - Determine and tune the configuration of all layers to ensure full throughput in and across each layer/service.
  - Measure and determine efficiency of the end-to-end solutions.
  - Monitor, identify and mitigate error conditions.



100G High Throughput Data Program

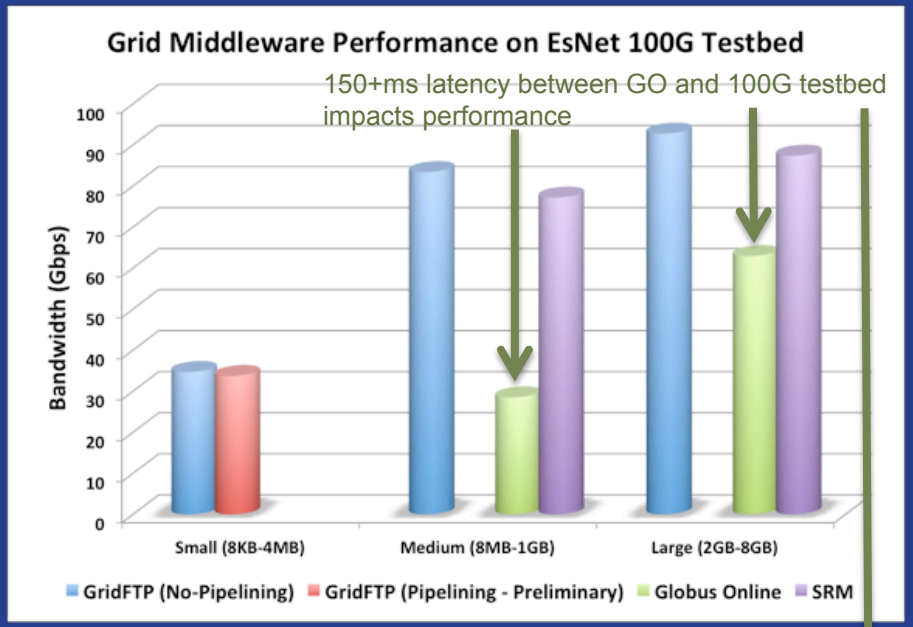
- 2011: Advanced Network Initiative (ANI) Long Island MAN (LIMAN) testbed.
  - GO / GridFTP over 3x10GE.
- 2011-2012: Super Computing '11
  - Fast access to ~30TB of CMS data in 1h from NERSC to ANL using GridFTP.
  - 15 srv / 28 clnt – 4 gFTP / core; 2 strms; TCP Win. 2MB

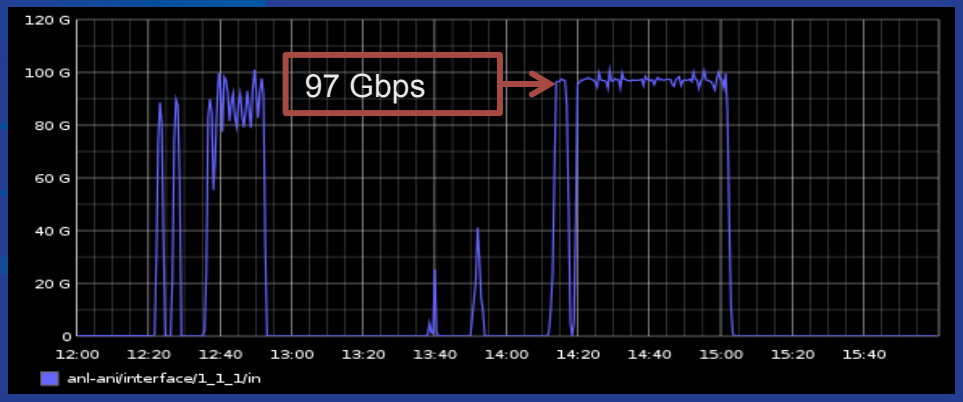


- 2012-2013: ESnet 100G testbed
  - Tuning parameters of middleware for data movement: xrootd, GridFTP, SRM, Globus Online, Squid. Achieved ~97Gbps
    - Rapid turn around on the testbed thanks to custom boot images
  - Commissioning Fermilab Network R&D facility: 8.5 Gbps per 10G node
- Spring/Summer 2013: 100G Endpoint at Fermilab
  - Validate hardware link w/ transfer apps for CMS current datasets
  - Test NFS v4 over 100G using dCache (collab. w/ IBM research)

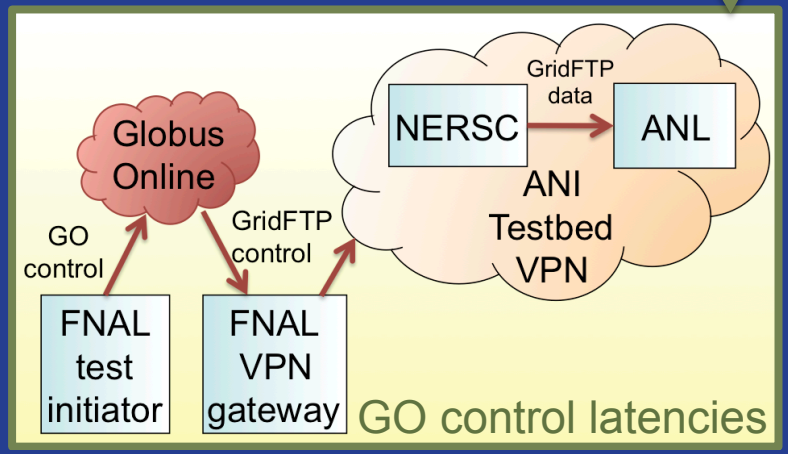
GridFTP / SRM / Globus Online Tests

- Data Movement using GridFTP
  - 3rd party Server to Server transfers: src at NERSC / dest at ANL
  - Dataset split into 3 size sets
- Large files transfer performance ~ 92Gbps
- Small files transfer – optimizing performance
- Issues uncovered on ESnet 100G Testbed:
  - GridFTP Pipelining lacks support for list of files & supports directory transfers only





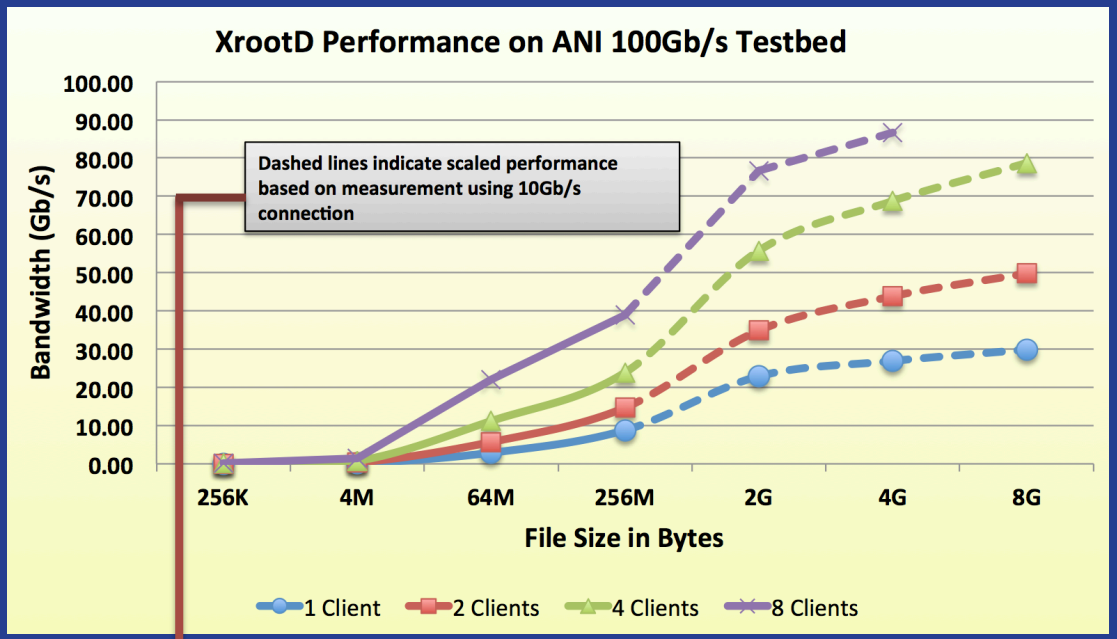
Optimal performance: 97 Gbps w/ GridFTP
2 GB files – 3 nodes x 16 streams / node



GO control channel sent to the VPN through port forwarding

XRootD Tests

- Data Movement over XRootD, testing LHC experiment (CMS / Atlas) analysis use cases.
  - Clients at NERSC / Servers at ANL
  - Using RAMDisk as storage area on the server side
- Challenges
  - Tests limited by the size of RAMDisk
  - Little control over xrootd client / server tuning parameters

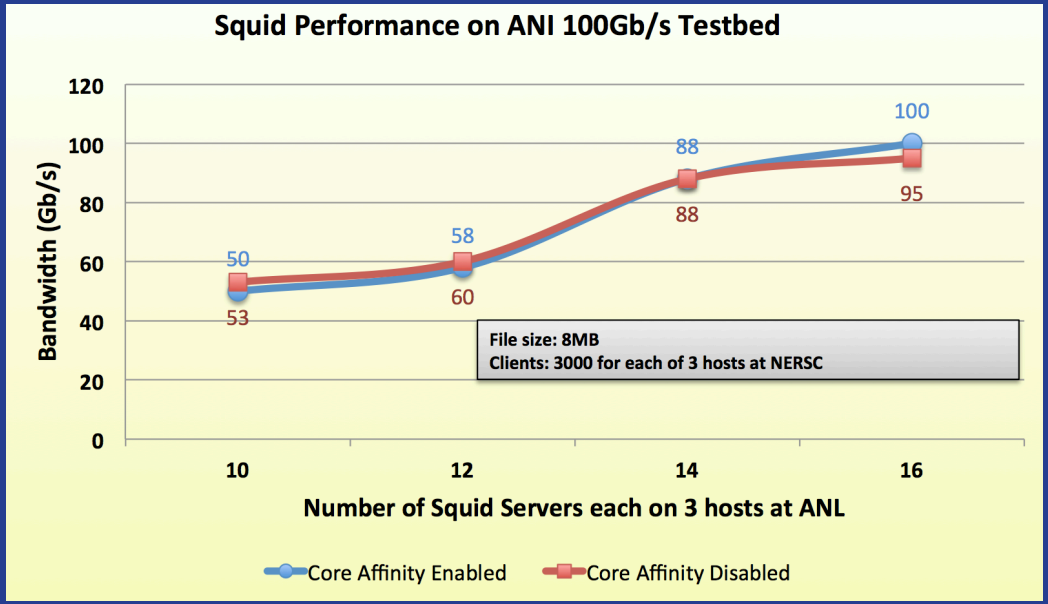


Dataset (GB)	1 NIC measurements (Gb/s)	Aggregate Measurements (12 NIC) (Gb/s)	Scale Factor per NIC	Aggregate estimate (12 NIC) (Gb/s)
0.512	4.5	46.9	0.87	–
1	6.2	62.4	0.83	–
4	8.7 (8 clients)	–	0.83	86.7
8	7.9 (4 clients)	–	0.83	78.7

Calculation of the scaling factor between 1 NIC and an aggregated 12 NIC for datasets too large to fit on the RAM disk

Squid / Frontier Tests

- Data transfers
  - Cache 8 MB file on Squid – This size mimics LHC use case for large calib. data
  - Clients (wget) at NERSC / Servers at ANL
  - Data always in RAM
- Setup
  - Using Squid2: single threaded
  - Multiple squid processes per node (4 NIC per node)
  - Testing core affinity on/off: pin Squid to core i.e. to L2 cache
  - Testing all clients v/s all servers AND aggregate one node v/s only one server



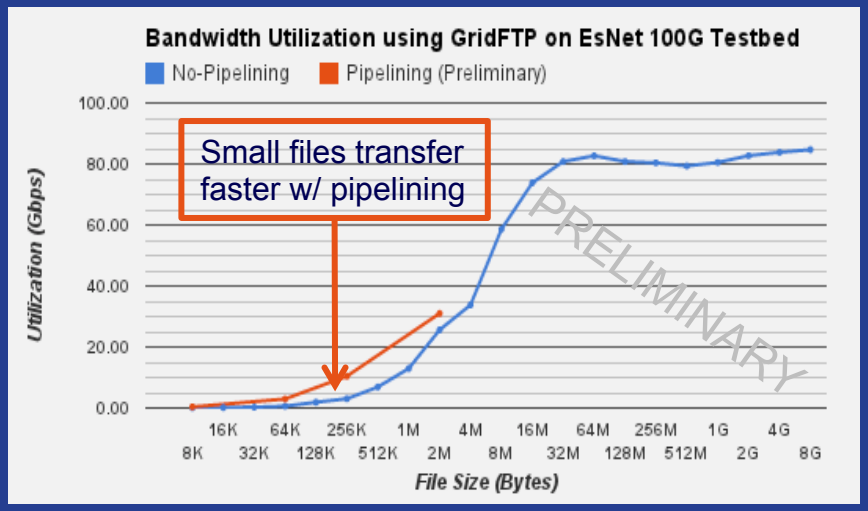
- Results
  - Core-affinity improves performance by 21% in some tests
  - Increasing the number of squid processes improves performance
  - Best performance w/ 9000 clients: ~100 Gbps

U.S. DEPARTMENT OF ENERGY

Currently Working On…

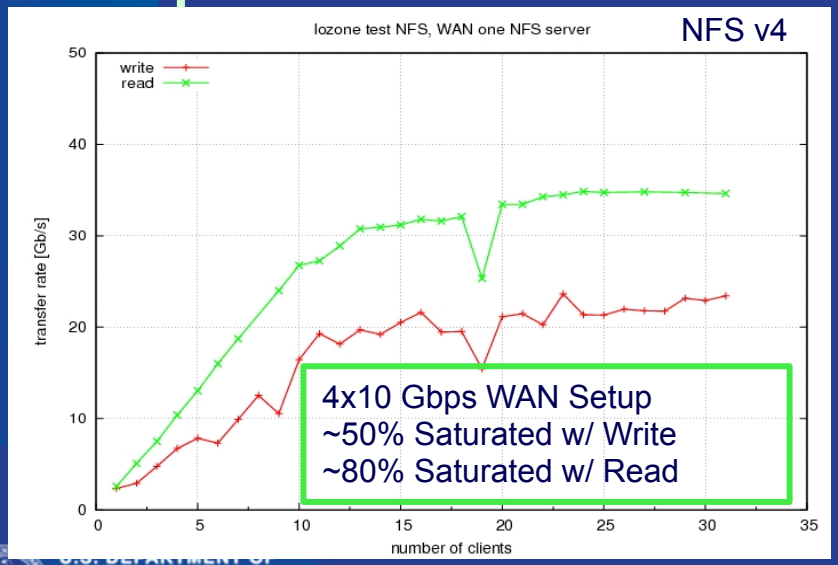
GridFTP Small Files

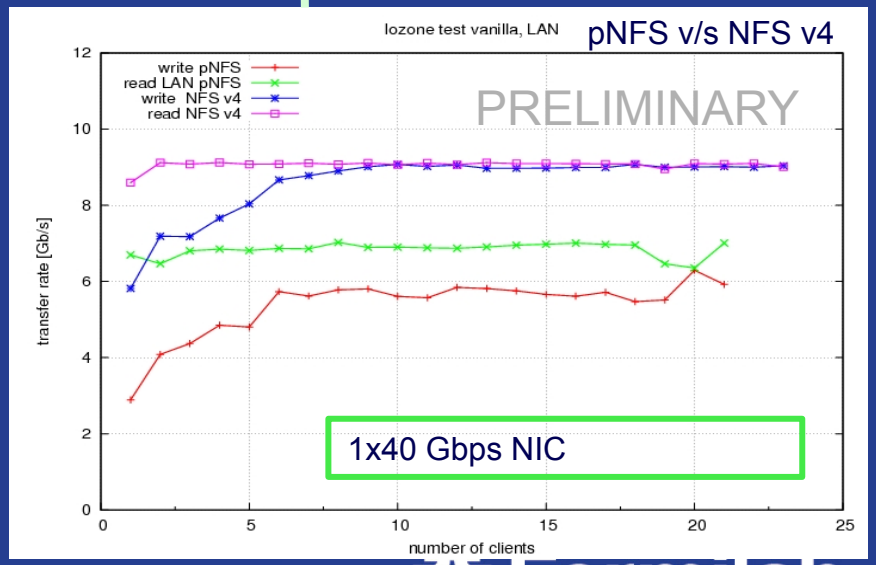
- Optimizing transfers varying pipelining depth & concurrency
- Comparing bandwidth utilization w/ and w/o pipelining.
- **Issues:** Pipelining interferes with concurrency



NFS v4 & pNFS

- Collaboration with IBM Research
- Mounting remote disks using NFS over 100G
- Validating dCache implementation of NFS v4





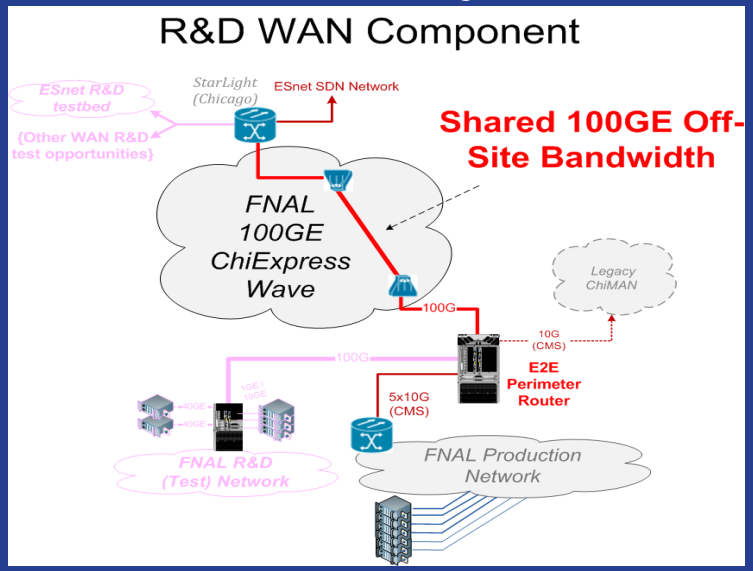
NFS Plots Courtesy of Dmitry Litvintsev

FermiCloud 100G Testbed
A dedicated R&D Network facility

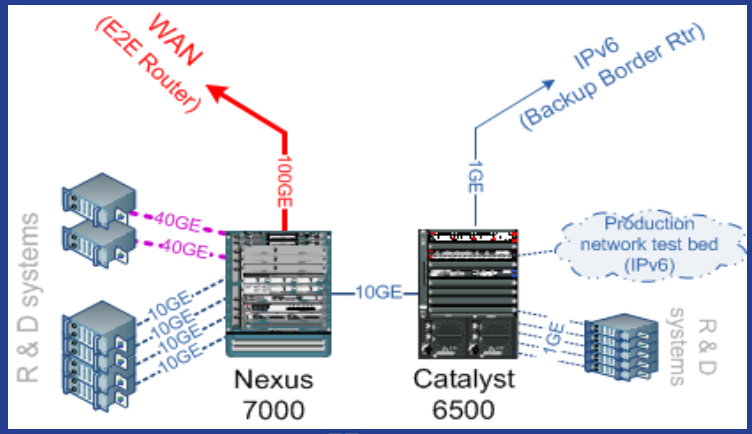
- 100G wave will support 50G of CMS traffic
- Remaining ~50G for FNAL R&D network
  - Potentially higher when CMS traffic levels are low
- Planning WAN circuit into ESnet 100G testbed
  - Potential for circuits to other R&D collaborations

- **100G R&D**
- **Production-like environment for tech evaluation**
- **Testing of firmware upgrades**
- **Work closely with Networking Group in CCD**

- Nexus 7000 w/ 2-port 100GE module / 6-port 40GE module / 10GE copper module
- 12 nodes w/ 10GE Intel X540-AT2 (PCIe) / 8 cores / 16 GB RAM
- 2 nodes w/ 40GE Mellanox ConnectX®-2 (PCIe-3) / 8 cores w/ Nvidia M2070 GPU
- Catalyst 6509E for 1GE systems
  - IPv6 tests / F5 load balancer / Infoblox DNS, Palo Alto firewall



*** Diagram courtesy of Phil Demar**



100G Testing with External Collaborators

- 100G tests between Fermilab & UFL
  - End Goal: Demonstrate transfers on 100G scale
    - Ideally: Disk-to-disk 100G tests
    - Realistic: Memory-to-memory or Disk-to-Memory
      - Lack of High Performance Storage connected to 100G FermiCloud Testbed
  - Test Setup:
    - 10G x 12 servers on both sides
    - FNAL Endpoint:
      - FermiCloud 100G Testbed functional by end of July 2013
    - UFL Endpoint: Testbed to be functional in July 2013
- 100G tests between UCSD & Fermilab
  - Timeline: Fall 2013

CortexNet

- Idea: Group of services with access to network usage pattern & statistics working together to provide information about best possible route to transfer data
  - Proposal Submitted to ASCR
    - Collaboration with Professor Ioan Raicu from IIT
  - Depends on Active & Passive Network Monitoring
    - Feasibility study & Prototype development using GridFTP
      - Gather transfer statistics from GridFTP and upload them to PerfSONAR
    - Extend the approach using transfer statistics from XRootD
      - Planned for future & contingent to funding

OSG Networking Dashboard

- OSG Networking Dashboard
  - Network of PerfSONAR servers
  - Network usage statistics in a central repo
    - Human consumption: Plots/Graphs/Monitoring
    - Services consumption: Programmatic access
- HTDP's involvement in OSG Networking Dashboard
  - AuthN/Z Working Group
    - AuthN/Z infrastructure
    - Dashboard
      - Inter-component communication
      - User/Admin communication
  - User API/Use Cases Working Group
    - User communities
    - Use cases
    - Missing functionality
    - Possible scalability issues

HTDP Milestones

Milestones	Date
Super Computing 2011 - Demo	11/14/2011
Super Computing 2012 – Poster + Abstract	11/10/2012
CHEP 2012 – Poster Presentation	05/21/2012
CHEP 2012 – Journal Paper Submission	06/22/2012
CCGrid 2013 – Paper Submission	03/15/2013
CortexNet Proposal Submission Collaboration: Ioan Raicu from IIT	04/19/2013
CCGrid 2013 – Paper Presentation	05/14/2013
NFS/dCache + HTDP Paper	2013/2014

Additional Information

- Project Home
  - https://cdcvs.fnal.gov/redmine/projects/htdp/wiki
  - Mailing List: htdp@fnal.gov
- Weekly meetings
  - Thurs 11:00 am - Noon @ BSS/Req Room (WH4)